Finding New Human Minisatellite Sequences in the Vicinity of Long CA-Rich Sequences

Fabienne Giraudeau,[1] Elisabeth Petit,[1] Hervé Avet-Loiseau,[2] Yolande Hauck,[1] Gilles Vergnaud,[1,3,5] and Valérie Amarger[1,4]

[1]*Institut de Génétique et Microbiologie, Université Paris-Sud, 91405 Orsay CEDEX, France;* [2]*Cytogénétique et Hématologie, Institut de Biologie des Hôpitaux de Nantes, 44035 Nantes CEDEX, France;* [3]*Laboratoire de Génétique Moléculaire, Centre d'Etudes du Bouchet, 91710 Vert le Petit, France*

Microsatellites and minisatellites are two classes of tandem repeat sequences differing in their size, mutation processes, and chromosomal distribution. The boundary between the two classes is not defined. We have developed a convenient, hybridization-based human library screening procedure able to detect long CA-rich sequences. Analysis of cosmid clones derived from a chromosome 1 library show that cross-hybridizing sequences tested are imperfect CA-rich sequences, some of them showing a minisatellite organization. All but one of the 13 positive chromosome 1 clones studied are localized in chromosomal bands to which minisatellites have previously been assigned, such as the 1pter cluster. To test the applicability of the procedure to minisatellite detection on a larger scale, we then used a large-insert whole-genome PAC library. Altogether, 22 new minisatellites have been identified in positive PAC and cosmid clones and 20 of them are telomeric. Among the 42 positive PAC clones localized within the human genome by FISH and/or linkage analysis, 25 (60%) are assigned to a terminal band of the karyotype, 4 (9%) are juxtacentromeric, and 13 (31%) are interstitial. The localization of at least two of the interstitial PAC clones corresponds to previously characterized minisatellite-containing regions and/or ancestrally telomeric bands, in agreement with this minisatellite-like distribution. The data obtained are in close agreement with the parallel investigation of human genome sequence data and suggest that long human (CA)s are imperfect CA repeats belonging to the minisatellite class of sequences. This approach provides a new tool to efficiently target genomic clones originating from subtelomeric domains, from which minisatellite sequences can readily be obtained.

[The sequence data described in this paper have been submitted to the EMBL data library under accession nos. AJ000377–AJ000383.]

Tandem repeats represent an important proportion of vertebrate genomes and have been classified as satellites, midisatellites, minisatellites, and microsatellites according to the overall length of the entire array. In higher vertebrates, $(CA)_n$ microsatellites are the most numerous, with an average distance between two microsatellites of ~25 kb (Stallings et al. 1991). Ninety percent of human $(CA)_n$ microsatellite arrays are <40 bp and <1%–2% are longer than 30 repeats (Weber 1990). Minisatellites repeat units are usually 10–100 nucleotides long, and the array spans 0.5–100 kb. Chromosomal distribution of minisatellites in the human genome is highly skewed towards telomeres and ancestrally telomeric regions (Amarger et al. 1998).

The initial classification of minisatellites and microsatellites has now been strengthened on biological grounds by the demonstration that different modes of evolution operate on these two types of structures. Microsatellites mutate by replication slippage processes because of mispairing between the two strands during replication. They are stabilized by variant repeats, whose presence facilitates detection of the slipped strand DNA by the mismatch repair system (Strand et al. 1993; Heale and Petes 1995). Minisatellites mutate predominantly in the germ line (Jeffreys and Neumann 1997) through mechanisms, including gene conversion-like events, presumably arising from DNA double-strand breaks (DSBs), insensitive to internal variations within the tandem array (Buard and Vergnaud 1994; Jeffreys et al. 1994). However, a number of intermediate situations raise the question of the border between the mini- and microsatellite classes. For instance, mutation rates at some minisatellites including MS1 (D1S7) are sensitive to mismatch repair deficiencies (Hoff-Olsen et al. 1995) reminiscent of a microsatellite behavior. At the other end of the spectrum, some human $(CA)_n$ repeats have extremely long alleles, with internal heterogeneity (Wilkie and Higgs 1992). Also, the origin of both classes of tandem repeats is still poorly understood. Microsatellite arrays may arise by replication errors or as a result of nonhomologous end-joining repair following DNA DSB events (Liang et al. 1998), which can create de novo $(CA)_n > 20$ stretches.

[4]*Present address: Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, 753 24 Uppsala, Sweden.*
[5]Corresponding author.
E-MAIL Vergnaud@igmors.u-psud.fr; FAX 33 1 69 15 66 78.

Unequal crossing-over or replication slippage between fortuitous short-direct repeats have been invoked to provide the initial duplication event of some minisatellites in human and yeast (Haber and Louis 1998).

To better understand the nature and origin of surprisingly long human (CA)s, we developed a technology that efficiently identifies clones containing long CA-rich sequences by a simple hybridization procedure. This approach was applied to human cosmid and PAC genomic libraries. The analysis of a subset of sequences strongly supports the conclusion that most long human CA-rich sequences are imperfect. The genome distribution of positive clones is highly skewed towards telomeres and minisatellites can usually be found in the vicinity. This observation is further strengthened by the parallel investigation of the currently available chromosome 7 human sequence data. Twenty-two new minisatellites have here been successfully identified, establishing the validity of this approach of minisatellite cloning by vicinity with long (CA)s.

RESULTS

Identification of Probes Appropriate for the Identification of Long CA Arrays

Chromosome 1 Cosmid Library Screening and Sequence Analysis of Some Positive Clones

Five different $(CA)_n$-derived DNA sequences were tested for their ability to detect genomic clones containing long (>100 bp) perfect or imperfect (CA)s, rather than short $(CA)_n < 40$ microsatellites: (1) a long perfect synthetic $(CA)_n$ array; (2 and 3) two long natural imperfect (CA)s, R62 and R85, characterized previously in a search for rat minisatellite and microsatellite sequences (Amarger et al. 1998; Giraudeau et al. 1999); and (4 and 5) two synthetic imperfect (CA)s, 16C46 and 14C32.

The long perfect $(CA)_n$ probe strongly detects ~4% of the 20,000 human cosmid clones assayed. The sequencing of four fragments cross-hybridizing with the long perfect $(CA)_n$ probe reveals microsatellites with 20 repeats, showing that this probe is not efficient for the selective identification of the longer (CA)s (data not shown). Probes R62, R85, 14C32, and 16C46 give a signal above background on 0.4%–0.6% of the clones. Clones are often detected by more than one probe as shown in Table 1 for eight R62 positive clones. Six are also positive with at least one of the other probes. The fragments detected by the CA-rich probes (R62, R85, 16C46, 14C32) were analyzed further. The sequences responsible for the cross-hybridization are very CA-rich but none is a perfect $(CA)_n$ array. In five cases, a repeating unit ranging in size from 5 bp to 23 bp is observed (Table 1). An important variability between the different motifs along the array is seen because of either point mutations, insertion/deletions, or changes in the number of repeats of an internal $(CA)_n$ array. First and last motifs of the array are usually difficult to delineate because the flanking sequences are also in many cases CA-rich sequences with variants. In the last two cases (within c112-N1332 and c112-P0688), no repeat unit can be defined. The cross-hybridizing sequence is a complex stretch of dinucleotide repeats, mainly (CA)s and (CT)s interspersed with variant repeats. Uninterrupted stretches of CA repeats are short, the maximum being six repeats in N1332 and four in P0688. Among these R62 positive fragments, three (CEB117, CEB118, and CEB121) show a typical minisatellite behavior by Southern blot hybridization.

Chromosomal Assignment of Positive Cosmid Clones

A total of 22 cosmids detected by one or more of the imperfect $(CA)_n$ probes from the chromosome 1 library (R62, R85, 14C32, and 16C46) were then assigned to a chromosomal band by FISH and/or linkage (Fig. 1, circles; Table 1). Thirteen (59%) are subtelomeric, seven (32%) are interstitial, and two (9%) are juxtacentromeric. Unexpectedly, nine clones (five of which are located in a terminal band) do not originate from chromosome 1. Seven among the 13 chromosome 1 cosmids are in the telomeric bands. All but one are localized on 1p36.3 region and the last one gives a signal by FISH hybridization at both ends of the chromosome. Among the nontelomeric cosmid clones, two are localized on 1p34.35, one in 1p12, one in 1q42, and two others in a juxtacentromeric region.

Application of the Methodology to the Screening of a Total Human Genome PAC Library

Probe R62 detects clones with a very good signal-to-background ratio and will not detect a $(CA)_{22}$ array [but would still detect a longer $(CA)_{40}$ array, independently characterized from a pig cosmid library; data not shown]. R62 was thus selected to hybridize a high-density filter carrying ~20,000 independent PAC clones, corresponding to one human genome equivalent. The 42 clones giving the strongest signal were successfully assigned to a chromosomal band by FISH and/or linkage analysis and are represented by squares on the 550-band karyotype presented Figure 1. Twenty-five PACs are assigned to a terminal band (60%), 4 are juxtacentromeric (9.5%), and 13 are interstitial (30.5%).

Identification of Minisatellites Within Positive PAC and Cosmid Clones

The cosmid and PAC clones identified by R62 screening were searched for minisatellites as described in Amarger et al. (1998). The DNA from each clone was digested separately with three combinations of two restriction enzymes: *Alu*I and *Hae*III; *Alu*I and *Hin*fI;

Table 1. Description of DNA Sequences Cross-Hybridizing with (CA)-Rich Sequences

Cosmid[a]	Chromosomal localization[b]	Hybridization signal[c] R62	R85	16C46	14C32	Hybridizing sequence[d]	Length *Hae*III (kb)[e]	Minisatellite name[f]
c112-N1332 (AJ000377)	Chr1p34.35 (F)	+	−	+	−	stretches of $(CA)_{1-6}$ and $(CT)_{1-4}$ interspersed by variants with a variation CCC occuring periodically	1	
c112-J06107 (AJ000383)	Chr1pter (F, L)	+	+	−	+	CC V $CA(CA)_{1-2}G$ G	0.6	CEB121
C112-J1081 (AJ000382)	Chr2q13 (F)	+	−	−	−	A G V $GGTG(CA)_{1-3}TCACAGCCA$ AAAC C TT	0.5	
c112-J2362 (AJ000380)	Chr5pter (F, L)	+	−	+	+	C V $CGCACG(CA)_{1-4}$ G	1.5	
c112-I0724 (AJ000378)	Chr8qter (F, L)	+	+	+	+	G V $CA(CA)_{1-4}GCACGC$ T TG T T G	1.2	CEB118
c112-L1183 (AJ000381)	Chr9qter (F, L)	+	−	−	+	T CTCA A V $C(CA)_{1-2}GCCCA(CA)A_{1-2}CTCA$ T	0.3	
c112-M1148	Chr10qter (L)	+	−	+	+	not done	5.5	CEB117
c112-P0688 (AJ000379)	Chr16q21.22 (F)	+	−	−	−	sequence formed mainly by stretches of $(CA)_{1-4}$, $(CT)_{1-2}$	0.5	

[a]Cosmid name (EMBL accession no.).
[b]Chromosomal localization [as determined by FISH (F) and/or linkage analysis (L)].
[c]Hybridization signal with the different probes.
[d]Short description of the cross-hybridizing fragment. When possible, a consensus motif with sequence variants (point mutations, insertion/deletion variants) is indicated. Variants are found independently of each other.
[e]Length of the subcloned fragment.
[f]Minisatellite name.

*Hae*III and *Hin*fI. Seventy-three cosmid fragments with a size above 1.3 kb after the double digestion were excised from agarose and tested for the presence of a minisatellite by hybridization on a Southern blot. Three minisatellites (CEB117, CEB118, CEB119) were isolated (Table 2). Using the same approach, 316 PAC fragments were tested. Eighteen minisatellites derived from 15 independent PAC clones were identified. Their main characteristics (allele size, polymorphism) are presented in Table 2. Twenty out of the 22 new minisatellites identified are derived from the telomeric PAC or cosmid clones. One (or more) minisatellite was identified in half of the telomeric PAC clones. PAC 1 contains (at least) four minisatellites: UPS17, UPS21, UPS22 (Table 2), in addition to CEB 70, which was characterized previously and independently (Spurr et al. 1994). PAC 50 contains two minisatellites: UPS6 and UPS7.

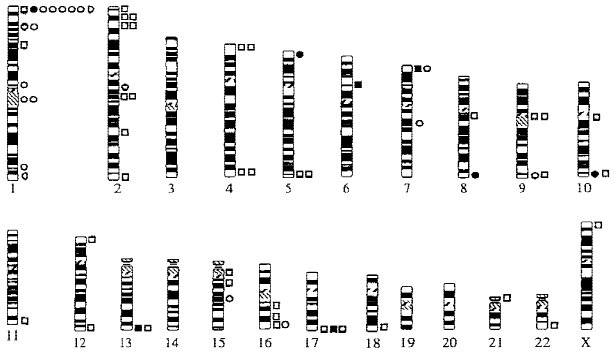

Figure 1 Chromosomal assignment of the clones detected in the human genome after hybridization of R62 probe on cosmid and PAC libraries. PAC (■) or cosmid (●) localized by FISH and linkage; PAC (□) or cosmid (○) localized by FISH or linkage. The two semicircles represent one cosmid clone revealing two locations by in situ hybridization.

Table 2. Description of Minisatellite Probes Identified from PACs and Cosmid Clones

PACs (RPCI6 no.)	Localization	Minisatellite(s) within	Fragment	Heterozygote frequency	No. of alleles	*Alu*I allele size (kb)
19 (213 J16)	2p23 (F)	UPS19	*Alu*I–*Hin*fI (1.7)	0/16	1	(1.8; 1.4; 1.3) allele cut
34 (196 L17)	4p16 (L) (ter)	UPS14	*Alu*I–*Hae*III (2.8)	7/16	3	4; 2.7; 2.4
28 (202 F12)	5p15.3 (L) (ter)	UPS9	*Alu*I–*Hae*III (2.1)	10/16	3	2.6; 2.4; 1.9
32 (208 C19)	6p21 (F, L)	UPS8	*Hae*III–*Hin*fI (2.1)	4/4	6	3.3; 2.4; 1.6; 1.65; 1.55; 1.3
21 (195 E8)	7p22 (F, L) (ter)	UPS5	*Hae*III–*Hin*fI (1.6)	3/4	3	1.9; 1.75; <0.5
33 (238 P2)	11q25 (L) (ter)	UPS3	*Alu*I–*Hae*III (1.7)	10/16	4	1.95; 1.90; 1.85; 1.7
6 (223 C10)	12p13.3 (F) (ter)	UPS15	*Alu*I–*Hae*III (1.5)	0/16	1	2.2
26 (207 H21)	12q24.33 (F) (ter)	UPS20	*Alu*I–*Hin*fI (1.7)	0/16	1	1.9
1 (217 O7)	13q34 (F, L) (ter)	UPS17	*Alu*I–*Hin*fI (2.4)	8/16	4	2.9; 2.85; 2.4; 1.8
		UPS21	*Alu*I–HinfI (1.3)	7/16	2	1.8; 1.3
		UPS22	*Alu*I–*Hin*fI (1.8)	5/6	2	1.8; 1.6
50 (210 M23)	13q34 (F, L) (ter)	UPS6	*Alu*I–*Hin*fI (1.6)	8/16	3	1.8; 1.6; 1.3
		UPS7	*Alu*I–*Hae*III (1.9)	13/16	6	2.25; 2.20; 2.15; 1.95; 1.75; 1.65
13 (224 C15)	17q25 (F, L) (ter)	UPS12	*Alu*I–*Hae*III (1.5)	2/16	3	1.65; 1.6; 1.55
25 (213 P13)	17q25 (L) (ter)	UPS4	*Hae*III–*Hin*fI (1.45)/ *Alu*I–*Hae*III (1.6)	10/16	4	1.95; 1.7; 1.65; 1.55
36 (238 N11)	21p13 (L) (ter)	UPS13	*Hae*III–*Hin*fI (2.9)/ *Hae*III–*Hin*fI (3.1)	3/4	7	4.3; 3.8; 3.0; 2.6; 2.55; 2.4; <0.5
49 (240 M9)	18q23 (L) (ter)	UPS1	*Alu*I–*Hae*III (2.4)/ *Hae*III–*Hin*fI (2.5)	15/16	7	4.2; 3.8; 3.75; 3.5; 3.45; 3.0; 2.9
24 (231 K15)	22q13 (F) (ter)	UPS11	*Alu*I–*Hin*fI (1.4)	6/16	2	1.8; 1.6
Cosmids		**Name**				***Hin*fI or *Hae*III allele size (kb)**
c112–J06107	1p36.6 (F, L) (ter)	CEB121	*Hae*III (0.6)	12/16	5	1; 1.4; 1.45; 1.75; 1.8 (*Hin*fI)
c112–J2362	5p15.3 (F, L) (ter)	CEB119	*Hae*III (1.5)	11/16	7	1; 1.2; 1.35; 1.4; 1.55; 1.8; 4.2 (*Hin*fI)
c112–I0724	8q24.3 (F, L) (ter)	CEB118	*Hae*III (1.2)	4/16	2	1.05 + 1; 0.95 + 1.8 (*Hae*III)
c112–M1148	10q26 (F, L) (ter)	CEB117	*Hae*III (0.5)	9/16	4	5.4; 5.5; 5.6; 6.5 (*Hin*fI)

In three cases, UPS15, UPS19, and UPS20, only one allele was detected. We assume that these sequences are tandem repeats because of the strong signal intensity obtained on Southern blots and because of the large allele size detected on *Alu*I, *Hae*III, and *Hin*fI digests (frequent cutters). (ter) Terminal band.

Parallel Investigation of Sequence Databases

The current status of publicly available human sequence data is reflected at http://www.ncbi.nlm.nih.gov/genome/seq/. Significant progress has already been achieved for a number of chromosomes, such as chromosomes 7, 17, 21, and 22, so that the screening of genome libraries can be compared to some extent to the direct screening of genome sequence. We selected chromosome 7 for further investigations, because the available sequence is relatively well distributed along the whole chromosome (i.e., in contrast with chromosome 17) and because the distribution of minisatellites on chromosome 7 has been well documented in earlier reports (Amarger et al. 1998). At the time of this investigation, ~54 Mb of sequence data was available, corresponding to 30% of a total estimate of 170 Mb for chromosome 7. Figure 2 presents some of the results obtained by searching and locating minisatellites and long CA sequences along the chromosome. As a reminder, Figure 2A (left) is compiled from this report

and Amarger et al. (1998) and locates minisatellite loci obtained by screening cosmid or PAC libraries. Figure 2B presents the density of tandem repeats with a repeat unit of 20 nucleotides or more, spanning at least 1000 nucleotides as identified in the sequence data using the tandem repeat finder described in Benson (1999). Figure 2C presents the relative density of long (spanning at least 300 nucleotides) CA-rich sequences detected by a FASTA search against the chromosome 7 sequence data using a 800 bp-long $(CA)_{400}$ as the query. None of the matches in this range is a perfect CA repeat. Four matches span >800 bp, two of which originate from 7q36 (no higher order organization of the degenerate CA rich array could be found; data not shown).

DISCUSSION

Minisatellites and microsatellites are two important classes of tandem repeats used as genome markers. Minisatellites have been shown to be useful tools to detect chromosomal rearrangements in a number of pathological situations, including mental retardation (Flint et al. 1995; Giraudeau et al. 1997). Some of them are suspected to be involved in gene regulation (Bennett et al. 1995). Very unstable human minisatellites have been characterized (Jeffreys et al. 1988; Vergnaud et al. 1991). The mutation rate is apparently increased by environmental agents such as radiation (Dubrova et al. 1997). To characterize new human minisatellites, as well as to investigate the boundary between mini- and microsatellites, we have devised a strategy enabling the rapid cloning of long CA-rich sequences from total genomic libraries by hybridization screening. Five CA-rich probes were evaluated for their ability to discriminate long CA-rich sequences from ordinary microsatellites. The long perfect $(CA)_n$ array does not discriminate against fragments containing a $(CA)_{20}$ array. The imperfect synthetic tandem repeat 16C46 [16-bp repeat unit containing an internal stretch of four (CA)s], is more appropriate but cross-hybridizes with a cosmid clone containing a $(CA)_{22}$ array (data not shown) and fails to detect many CA-rich sequences (Table 1). Probe R85, with internal stretches of up to three (CA)s, detects clones with weak intensity. Probe R62 appears to be a very good compromise. It will not detect a perfect $(CA)_{22}$, although it will detect a longer $(CA)_{40}$ stretch. It detects many clones, with a good signal-to-background ratio, as compared to the synthetic 14C32 array. In contrast with 14C32, R62 also detects complex stretches of imperfect (CA)s devoid of higher-order organization (Table 1).

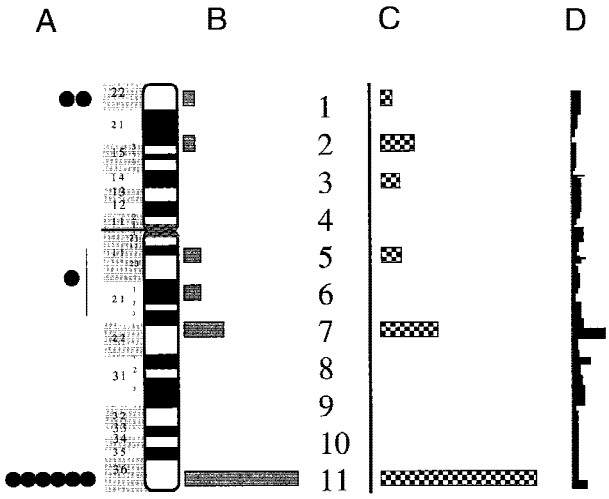

Figure 2 Comparison of genomic libraries and sequence database investigations for HSA Chr 7. (*A*) The position of minisatellite loci characterized in this report (one locus) and in Amarger et al. (1998) for chromosome 7, showing the predominantly, but not exclusively, telomeric location of minisatellites on this chromosome and the imbalance between the two chromosome ends. (*B*) The density of tandem repeats as detected in the sequence data (repeat unit >20 bp, repetition spanning at least 1000 nucleotides). (*C*) The density of long CA-rich sequences (in which the similarity, as detected by FASTA, with the 800-bp long CA_{400} query spans 300 bp or more). (*D*) The distribution of ESTs characterized so far along the chromosome. The numbering (1–11) refers to the eleven bins defined in the Methods section. The largest peak, in *B* and *C*, corresponds to a density of, respectively, two and three qualifying objects per megabase of sequence available for this bin.

First, to reveal the chromosomal distribution pattern of long CA sequences, a chromosome 1-specific cosmid library was screened. Among the 22 cosmids studied, 13 originate from chromosome 1 (Fig. 1). The nine cosmids coming from other chromosomes, including five telomeric loci, presumably reflect some contamination of the library. The fact that the actual contamination of the library is less than the proportion of non-chromosome 1 cosmids in our selection (40%) suggests a telomeric bias in the contamination of the chromosome 1 library. Seven of the 13 chromosome 1 cosmids are within the terminal 1p36.3 band, where a minisatellite cluster was previously and independently described (Amarger et al. 1998). Two are localized on 1p34.35 where minisatellite MS1 (D1S7) is localized. Two others are localized in a juxtacentromeric region where the *MUC1* gene characterized by tandem repeats units has been isolated. Another is localized in 1q42, containing minisatellite MS32 (D1S8). Overall, the chromosome 1 distribution pattern of long CA-rich sequences is highly similar to the chromosome 1 minisatellite distribution pattern that is shown in Amarger et al. (1998) or that can be deduced from the NIH/CEPH Collaborative Mapping Group (1992) data, suggesting that the procedure could be applied on a larger scale for the identification of minisatellite associated regions.

For this purpose, a whole-genome PAC library was screened using the R62 probe to enable the cloning of new minisatellite sequences in the vicinity of CA-rich sequences. A significant proportion (25/42) of the R62-positive PAC clones studied are assigned to a terminal

band of the karyotype and 13 of them contain new minisatellites. The juxtacentromeric location of four (9.5%) PAC clones may reflect a peculiar behavior of these regions or may indicate an ancestrally telomeric location. Thirteen clones (30.5%) are interstitial. One of them is assigned to band 2q13 (Fig. 1) which is the position of a well-characterized chromosome fusion site (IJdo et al. 1991). Another one is located on 1p31-p32, where one minisatellite was described previously (Amarger et al. 1998). This human chromosomal region is homologous with 6qter in pig (Amarger et al. 1998). As shown here in PAC 1 and PAC 50 (Table 2), the use of large insert clones further emphasizes the clustering of minisatellites within telomeric regions (Vergnaud et al. 1993).

The predominantly telomeric distribution is highly reminiscent of the distribution of minisatellites across the human genome and clearly different from the even distribution of microsatellites. In good agreement with this, a similar result is obtained by the investigation of sequence databases using a FASTA search (Pearson and Lipman 1988) and $(CA)_{400}$ as the query. Although long perfect (CA)s (>200 bp, e.g., accession no. Z81056 from *Caenorhabditis elegans*) are represented in the database, none of these originate from primates or even other mammals (data not shown). Figure 2C shows the density of hits spanning at least 300 bp, across human chromosome 7. All such hits are imperfect, CA-rich stretches, with or without higher-order redundancy. The distribution is almost identical to the patterns shown in Figure 2, A and B, which reflect chromosome 7 minisatellite distribution. No obvious correlation is seen between the chromosome 7 gene density presented in Figure 2D and the minisatellite and long (CA)s distribution, with the exception perhaps of segment 7, band 7q22, which is a common peak (Fig. 2).

METHODS

High-Density Filters from Human Genome Libraries

High-density filters corresponding to a human chromosome 1 cosmid library were obtained from the Max-Planck Institute for Molecular Genetics. This library is represented by two high-density filters with 20,000 clones spotted on each membrane. Each clone is named by the number (c112) of the library and a specific number.

High-density filters corresponding to a human PAC library were obtained from the Roswell Park Cancer Institute (RPCI) center (http://bacpac.med.buffalo.edu/; the RPCI6 segment was used).

Perfect and Scrambled $(CA)_n$ Arrays Probes

A perfect long $(CA)_n$ probe and the two imperfect $(CA)_n$ arrays 14C32 $(GACACACTCACAGC)_n$ and 16C46 $(CACACACATGCACATA)_n$ were synthesized as described in Vergnaud (1989). 14C32 and 16C46 were designed so as to contain a maximum of three and four uninterrupted CA repeats, respectively. The natural scrambled $(CA)_n$ arrays R62 (EMBL accession no. AC AJ000072) and R85 (EMBL accession no. AJ000073) were selected among rat minisatellite sequences (Pravenec et al. 1996; Amarger et al. 1998). R62 and R85 repeat units are $(CACACT)_{1-2}$CACAGYRR (14 or 20 bp) and $(CAGGACA)_{1-2}$ GTGARCACA (16 or 23 bp), respectively.

Probe Labeling and Hybridization

The DNA fragments were recovered from agarose by centrifugation through glass wool as described by Heery et al. (1990). The probes were labeled with [α-^{32}P]dCTP Institute of Chemical and Nuclear (ICN) by the random priming procedure (Feinberg and Vogelstein 1984). Hybridization was done as described in Vergnaud (1989) in an hybridization oven. After hybridization, the filters were washed in 1× SSC/0.1% SDS or 0.1× SSC/0.1% SDS. Hybridization and washing were done at 60°C (screening of library filters) or 65°C (hybridization of Southern blots).

Subcloning and Sequencing

Restriction digest fragments were recovered from agarose using the Jetsorb kit (Bioprobe System). The fragments were ligated into *Sma*I Puc 18 vector (Pharmacia) before transfer to *Escherichia coli* XL1 strain (Stratagene) by electroporation.

Recombinant plasmids were sequenced using ^{33}P-labeled direct and reverse M13 primers with the Delta *Taq* sequencing kit (U.S. Biochemical) in a Perkin Elmer GenAmp PCR System 9600 thermocycler.

Identification of Minisatellites Within PAC and Cosmid Clones

DNA from each PAC or cosmid clone was digested by *Alu*I and *Hae*III, *Alu*I and *Hin*fI, or *Hae*III and *Hin*fI. Fragments >1.3 kb in size were recovered from agarose and hybridized to Southern blots carrying two reference individuals digested separately by *Alu*I, *Hae*III, *Hin*fI, and *Pvu*II, as described in Amarger et al. (1998).

Chromosomal Assignment by Linkage Analysis

Linkage analysis was performed on the CEPH (Centre d'Etudes du Polymorphisme Humain) panel of human families. Genotypes were managed using GENBASE, developed by Jean Marc Sebaoun (Spurr et al. 1994). Linkage files output were converted to CRIMAP file format using the LINK2CRI utility software written by John Attwood. CRIMAP version 2.4 (Green et al. 1990) was used for the analyses.

Chromosomal Assignment by FISH

Cosmid or PAC DNAs were labeled with biotin by nick translation. After overnight hybridization on target chromosome spreads, slides were washed in 2× SSC at 37°C. Probes were detected with FITC-avidin and analyzed with an epifluorescence microscope (DMRB-Leica) equipped with a CCD camera driven by the Powergene system from Perceptive Scientific International (PSI).

Sequence Database Searches

Chromosome 7 sequence data (54 Mb available at the time of this investigation) were retrieved from the National Center for Biotechnology Information (NCBI) site at (http://www.ncbi.nlm.nih.gov/genome/seq/). Tandem repeats were identified using the online software accessible at http://c3.biomath.mssm.edu/trf.html (Benson 1999). Large CA-rich se-

quences were detected using FASTA (Pearson and Lipman 1988) and a $(CA)_{400}$ synthetic sequence as the query. The FASTA analysis was done using the computing facilities provided by Infobiogen (information at http://www.infobiogen.fr/). Each sequence contig was assigned to a bin along the chromosome. Eleven bins of equal size were defined. The horizontal bars presented in Figure 2 represent the density of the object category per megabase of sequence in the corresponding bin. The current distribution of human ESTs on chromosome 7 was retrieved from the NCBI site (http://www.ncbi.nlm.nih.gov/genemap/).

ACKNOWLEDGMENTS

We thank the Resource Center/Primary Database of the German Human Genome Project, Berlin, Germany for providing the human cosmid clones. We thank Olivier Raineteau and France Denoeud for their participation at different stages of this project as summer students. This work was supported by the EUROGEM project (EC contract GENE-CT93-0101), the PiGMaP project (VA; EC contract BIO2-CT94-3044), an Action Concertée Coordonée-Sciences de la Vie grant from the French Ministry of Research, and by a grant from La Ligue contre le Cancer (Département de Vendée, France) to F.G.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Amarger, V., D. Gauguier, M. Yerle, F. Apiou, P. Pinton, F. Giraudeau, S. Monfouilloux, M. Lathrop, B. Dutrillaux, J. Buard et al.. 1998. Analysis of the human, pig, and rat genomes supports a universal telomeric origin of minisatellite sequences. *Genomics* **52:** 62–71.

Bennett, S.T., A.M. Lucassen, S.C.L. Gough, E.E. Powell, D.E. Undlien, L.E. Pritchard, M.E. Merriman, Y. Kawaguchi, M.J. Dronsfield, F. Pociot et al. 1995. Susceptibility to human type 1 diabetes at IDDM2 is determined by tandem repeat variation at the insulin gene minisatellite locus. *Nat. Genet.* **9:** 284–292.

Benson, G. 1999. Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Res.* **27:** 573–580.

Buard, J. and G. Vergnaud. 1994. Complex recombination events at the hypermutable minisatellite CEB1 (D2S90). *EMBO J.* **13:** 3203–3210.

Dubrova, Y.E., V.N. Nesterov, N.G. Krouchinsky, V.A. Ostapenko, G. Vergnaud, F. Giraudeau, J. Buard, and A.J. Jeffreys. 1997. Further evidence for elevated human minisatellite mutation rate in Belarus eight years after the Chernobyl accident. *Mut. Res.* **381:** 267–278.

Feinberg, A.P. and B. Vogelstein. 1984. Addendum: A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity. *Anal. Biochem.* **137:** 266–267.

Flint, J., A.O.M. Wilkie, V. Buckle, R.M. Winter, A.J. Holland, and H.E. McDermid. 1995. The detection of subtelomeric chromosomal rearrangements in idiopathic mental retardation. *Nat. Genet.* **9:** 132–139.

Giraudeau, F., D. Aubert, I. Young, S. Horsley, S. Knight, L. Kearney, G. Vergnaud, and J. Flint. 1997. Molecular-cytogenetic detection of a deletion of 1p36.3 leads to a revised estimate of the frequency of subtelomeric rearrangements in idiopathic mental retardation. *J. Med. Genet.* **34:** 314–317.

Giraudeau, F., F. Apiou, V. Amarger, P.J. Kaisaki, M.T. Bihoreau, M. Lathrop, G. Vergnaud, and D. Gauguier. 1999. Linkage and physical mapping of rat microsatellites derived from minisatellite loci. *Mamm. Genome* **10:** 405–409.

Green, P., K. Falls, and S. Crooks. 1990. *Documentation for CRI-MAP, version 2.4.* Washington University School of Medicine, St. Louis, MO.

Haber, J.E. and E.J. Louis. 1998. Minisatellite origins in yeast and humans. *Genomics* **48:** 132–135.

Heale, S.M. and T.D. Petes. 1995. The stabilization of repetitive tracts of DNA by variant repeats requires a functional mismatch repair system. *Cell* **83:** 539–545.

Heery, D.M., F. Gannon, and R. Powell. 1990. A simple method for subcloning DNA fragments from gel slices. *Trends Genet.* **6:** 173.

Hoff-Olsen, P., G.I. Meling, and B. Olaisen. 1995. Somatic mutations in VNTR locus D1S7 in human colorectal carcinomas are associated with microsatellite instability. *Hum. Mutat.* **5:** 329–332.

IJdo, J.W., A. Baldini, D.C. Ward, S.T. Reeders, and R.A. Wells. 1991. Origin of human chromosome 2: An ancestral telomere-telomere fusion. *Proc. Natl. Acad. Sci.* **88:** 9051–9055.

Jeffreys, A.J. and R. Neumann. 1997. Somatic mutation processes at a human minisatellite. *Hum. Mol. Genet.* **6:** 129–136.

Jeffreys, A.J., N.J. Royle, V. Wilson, and Z. Wong. 1988. Spontaneous mutation rates to new length alleles at tandem-repetitive hypervariable loci in human DNA. *Nature* **332:** 278–281.

Jeffreys, A.J., K. Tamaki, A. MacLeod, D.G. Monckton, D.L. Neil, and J.A.L. Armour. 1994. Complex gene conversion events in germline mutation at human minisatellites. *Nat. Genet.* **6:** 136–145.

Lehrach, H. 1990. *Genome analysis: Genetic and physical mapping.* (ed. K.E. Davies and S.M. Tilghman), pp. 39–81. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Liang, F., M. Han, P.J. Romanienko, and M. Jasin. 1998. Homology-directed repair is a major double-strand break repair pathway in mammalian cells. *Proc. Natl. Acad. Sci.* **95:** 5172–5177.

NIH/CEPH Collaborative Mapping Group. 1992. A comprehensive genetic linkage map of the human genome. *Science* **258:** 67–83.

Pearson, W.R. and D.J. Lipman. 1988. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci.* **85:** 2444–2448.

Pravenec, M., D. Gauguier, J.-J. Schott, J. Buard, V. Kren, V. Bila, C. Szpirer, J. Szpirer, J.-M. Wang, H. Huang et al. 1996. A genetic linkage map of the rat derived from recombinant inbred strains. *Mamm. Genome* **7:** 117–127.

Spurr, N.K., S.P. Bryant, J. Attwood, K. Nyberg, S.A. Cox, A. Mills, R. Bains, D. Warne, L. Cullin, S. Povey et al. 1994. European Gene Mapping Project (EUROGEM): Genetic maps based on the CEPH reference families. *Eur. J. Hum. Genet.* **2:** 193–203.

Stallings, R.L., A.F. Ford, D. Nelson, D.C. Torney, C.E. Hildebrand, and R.K. Moyzis. 1991. Evolution and distribution of $(GT)_n$ repetitive sequences in mammalian genomes. *Genomics* **10:** 807–815.

Strand, M., T.A. Prolla, R.M. Liskay, and T.D. Petes. 1993. Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair. *Nature* **365:** 274–276.

Vergnaud, G. 1989. Polymers of random short oligonucleotides detect polymorphic loci in the human genome. *Nucleic Acids Res.* **17:** 7623–7630.

Vergnaud, G., D. Gauguier, J.-J. Schott, D. Lepetit, V. Lauthier, D. Mariat, and J. Buard. 1993. Detection, cloning, and distribution of minisatellites in some mammalian genomes. In *DNA fingerprinting: State of the science.* (ed. S.D.J. Pena, R. Chakraborty, J.T. Epplen, and A.J. Jeffreys), Vol. 67, pp. 47–57. Birkhäuser Verlag, Basel, Switzerland.

Vergnaud, G., D. Mariat, F. Apiou, A. Aurias, M. Lathrop, and V. Lauthier. 1991. The use of synthetic tandem repeats to isolate new VNTR loci: Cloning of a human hypermutable sequence. *Genomics* **11:** 135–144.

Weber, J.L. 1990. Informativeness of human $(dC\text{-}dA)_n$ $(dG\text{-}dT)_n$ polymorphisms. *Genomics* **7:** 524–530.

Wilkie, A.O. and D. Higgs. 1992. An unusually large $(CA)_n$ repeat in the region of divergence between subtelomeric alleles of human chromosome 16p. *Genomics* **13:** 81–88.

Received March 1, 1999; accepted in revised form May 25, 1999.